Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 2012 Proceedings | Proceedings

The State of Social Computing Research: A Literature Review and Synthesis using the Latent Semantic Analysis Approach

Yibai Li
Department of Management, Information Systems, and Entrepreneurship (MISE), Washiongton State University, Pullman, WA, United States., yibaili@hotmail.com

K.D. Joshi
Department of Management, Information Systems, and Entrepreneurship (MISE), Washington State University, Pullman, WA, United States., joshi@wsu.edu

The State of Social Computing Research: A Literature Review and Synthesis using the Latent Semantic Analysis Approach

Yibai Li
Department of Management, Information Systems, and Entrepreneurship (MISE)
College of Business
Washington State University
yibai@wsu.edu

K.D. Joshi
Department of Management, Information Systems, and Entrepreneurship (MISE)
College of Business
Washington State University
joshi@wsu.edu

ABSTRACT

Social computing is an emerging research discipline. The number of publications on social computing has increased by 120% annually in the past four years. Despite the proliferation of studies in this area there is a lack of comprehensive, unified, and systematic characterization of this phenomenon. The definition and characterization of this phenomenon in the extant literature is diverse and fragmented. In this paper we attempt to bring some clarity by synthesizing and summarizing the extant literature in this area. We use Latent Semantic Analysis (LSA), a text mining and natural language processing technique, to summarize the state of social computing research. The results show that there are 27 unique dimensions which currently characterize this concept. LSA also reveals that, the 266 articles found in the literature predominantly focus on three major research themes namely, Knowledge Discovery, Knowledge Sharing, and Content Management in the Social Computing context.

Keywords

Social Computing, Social Computing Research, Latent Semantic Analysis, LSA, Literature Review, Text Mining, Natural Language Processing

INTRODUCTION

Although the concept of social computing can be traced back to the 1940s in Vannevar Bush's seminal 1945 Atlantic Monthly paper "As We May Think." (Wang Carley Zeng and Mao 2007), the explosive use of this term in the literature starts in the 1990s (Figure 1). As depicted in Figure 1, which is derived from our literature review data, the Social Computing research experienced a significant increase in 2007 and maintained a 120% growth rate annually since then. It has become a hot topic attracting interest from not only researchers but also technologists, software and online game vendors, Web entrepreneurs, business strategists, political analysts, and digital government practitioners (Wang et al. 2007). Noticing this emerging research area, many journals such as the ACM of Communication in 1994 January, IEEE Intelligent Systems in 2007 March, Decision Science in 2012 and IEEE Internet Computing in 2010 released special issues on social computing research. IEEE Computer Society publishes a special bi-monthly column on the topic of social computing (John 2011).

It is believed that Social Computing (SC) represents the new phase on the web (Parameswaran and Whinston 2007a). As the broadband connectivity and powerful personal computing devices becomes readily available to individual users on the internet, social computing is expected to empower individual users and eventually mitigate the information asymmetry by broadening the information flow (Parameswaran and Whinston 2007b). Some of the Social Computing initiatives have led to real business models such as blogging; Wikipedia; flickr; social networks like orkut, MySpace, Bebo, FaceBook, and LinkedIn. However, despite the proliferation of Social Computing in practice, systematically studying and researching this phenomenon can be challenging due to its rapid growth and fast changing nature. Despite the recent publication growth in this area (or possible due to this growth), there is a lack of comprehensive, unified, and systematic characterization of this phenomenon. The current characterization of this phenomenon in the business and scholarly literature is diverse and fragmented. In the absence of a comprehensive and systematic characterization of a research field, a field's "*progress is but a fortunate combination of circumstances, research is fumbling in the dark, and dissemination of knowledge is a cumbersome process*"(Vatter 1947). ***Therefore, in this paper we attempt to bring some clarity to the field of Social Computing by synthesizing and summarizing the extant literature in this area. More specifically we explore the following two questions***

in the paper: (1) How is Social Computing defined in various research studies? and (2) what constitutes Social Computing research?

To answer the aforementioned questions, this study conducted a Latent Semantic Analysis (LSA) on the existing Social Computing literature. The results show that Social Computing indeed is an ill-defined concept that is viewed and characterized differently by various researchers. Our analysis indicates that social computing has 27 components each representing a unique aspect of this phenomenon. In terms of the research themes in SC, our analysis shows that the 266 articles published on SC predominantly focus on three themes: Knowledge Discovery, Knowledge Sharing and Content Management. A ten-factor solution uncovered by LSA reveals that current research in this area mainly converges into the aforementioned three major research themes. The remainder of this article is organized as follows: we first introduce the Latent Semantic Analysis (LSA) and describe the application of this approach in this study. In the results section we analyze the statistics derived from LSA and address the posited research questions.

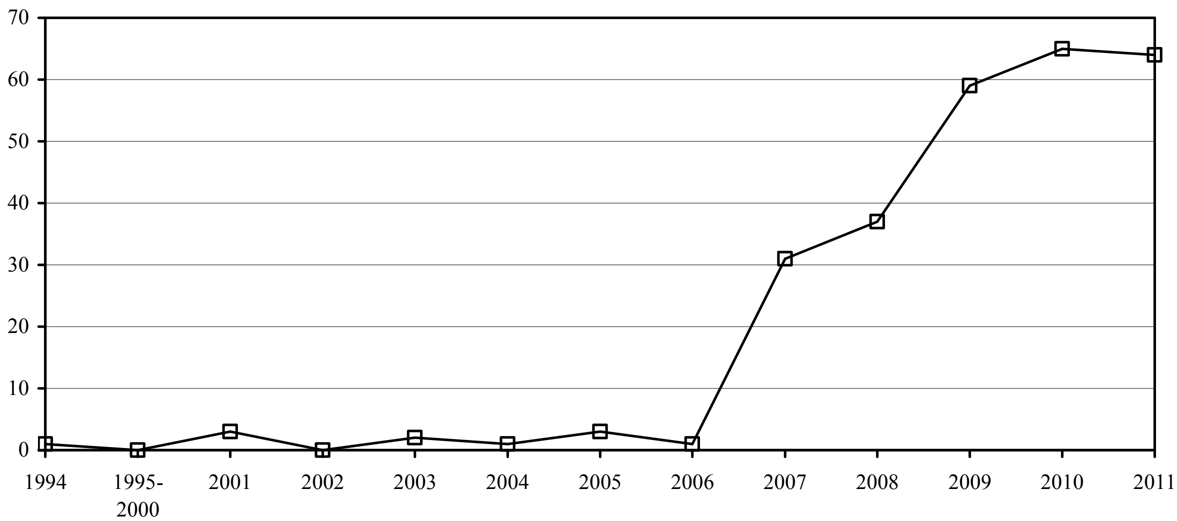

Figure 1. The Number of Publications on Social Computing From 1994 To 2011

METHOD

Data Collection

In order to characterize the Social Computing research landscape, this study searched for all available papers that contain the phrase "social computing" in their abstracts or author supplied key words from Business Source Complete (EBSCO), IEEE Xplore (IEEE), ACM Digital Library, and INFORMS PubsOnline (INFORMS) at the time of December 2011. The search result was restricted to those academic papers that contain the entire term -- "social computing" -- rather than just a single word "social" or "computing." We found 31 papers in EBSCO, 141 papers in IEEE, 171 papers in ACM Digital Library and 2 papers in INFORMS. After removing the duplicates, our sample comprised of 266 unique papers. We create two datasets from these 266 papers. First, we extracted all the Social Computing definitions existing in these 266 papers. 29 different definitions were identified. These definitions were consolidated in a spreadsheet which was used as the definition data set. The second dataset contained all the abstracts from the 266 papers. These two datasets were used to conduct the data analysis using the Latent Semantic Analysis approach.

Data Analysis

These two data sets were analyzed with Latent Semantic Analysis (LSA) using one of the leading data mining tools – the Rapidminer 5.0 (Jungermann 2009). Latent Semantic Analysis is a type of well-accepted text mining technique (Han Kamber and Pei 2011). Text mining is an umbrella term defined as "… the machine supported analysis of text. It uses techniques from information retrieval, information extraction as well as natural language processing (NLP) and connects them with the algorithms and methods of data mining, machine learning and statistics."(Hotho Nürnberger and Paaß 2005). It is also defined as "a process that employs a set of algorithms for converting unstructured text into structured data objects, and the quantitative method that analyze these data objects to discover knowledge" (Delen and Crossland 2008). In recent years, text mining has been increasingly used for knowledge discovery from scholarly literature (Delen et al. 2008; Jensen Saric and Bork 2006a; Mei and Zhai 2005; Turban Sharda Aronson and King 2008). The application of text mining is especially fruitful in biologic and genetic research. Because the number of articles published in those fields is increasing so quickly it is increasingly infeasible for a researcher to keep up-to-date with all of the relevant literature manually, even on specialized topics (Jensen Saric and Bork 2006b). Geneticists even use text mining to identify the complicated linkages between genes and diseases (Hristovski Peterlin Mitchell and Humphrey 2005). In summary, text mining empowers researchers to discover knowledge from a large amount of literature in a quantitative manner without researchers' bias.

A Brief Introduction to Latent Semantic Analysis

Among all kinds of text mining techniques, Latent Semantic Analysis (LSA) is a special mathematical and statistical method used to identify the latent concepts within the textual data at the semantic level (Hossain Prybutok and Evangelopoulos 2011). In contrast to many other text mining techniques which analyze textual data at the syntax level by simply counting the occurrence of particular words, LSA is a methodology that can extract the contextual-usage meaning of words and obtain approximate estimates of meaning similarities among words within the given textual data, thus providing the information at the semantic level (Hossain et al. 2011). LSA has numerous applications in natural language processing, search engine and library indexing and many other areas (Hossain et al. 2011).

LSA simulates the way the human brain distills meaning from text (Sidorova Evangelopoulos Valacich and Ramakrishnan 2008). LSA is capable of identifying underlying concepts within textual data for its particular mathematical method considers not only the word frequency per se but also the contexts in which the particular word is embedded (Sidorova et al. 2008). LSA is based on the fact that multiple words may share the same meaning and one word may mean different things in different contexts. The words that share the same meaning will "load" to their common underlying concept; one word may "load" to multiple latent concepts other than its main underlying concept.

LSA generates two sets of loadings, one for the terms (or words) and one for the documents (the definitions or abstracts in this study). The term loading shows how individual terms or words load to different latent concepts. Higher term loading reflects the greater chance that the particular term is truly associated with a certain latent concept. Likewise, the document loadings shows how different documents load to different latent concepts. Higher document loading means a greater likelihood that the particular document is truly talking about a certain latent concept. LSA also generates a singular value matrix (the square roots of eigenvalues) which shows the importance of all identified latent concepts. A higher singular value is associated with a greater importance of particular latent concepts. The researchers can use their judgment to choose the cut-off point for the eigenvalue, a point below which a latent concept is too "trivial" to be considered in the study. Researchers can adjust the cut-off point to get different level of the aggregation. At a lower level of aggregation, factors will reveal common research themes and, at a higher level of aggregation, key research areas. The detailed mathematical explanation for LSA can be found in previous studies (Sidorova et al. 2008).

The Operationalization of LSA in this study

This study follows the well-established text mining procedures as discussed in prior studies (Delen et al. 2008; Fox 1992; Han et al. 2011; Harman 1992; Hossain et al. 2011; Sidorova et al. 2008; Turban et al. 2008). A total of 266 abstracts from all existing articles on Social Computing were consolidated in a spreadsheet. The two data sets were respectively loaded to Rapidminer 5.0 and were processed through text mining procedures and matrix operations: term reduction, term frequency matrix transformation, and Singular Value Decomposition.

Pre-Processing and Term Reduction: First, the spreadsheet was converted into a document object in Rapidminer 5.0 and was assigned a unique document ID before it can be analyzed. Then the documents went through a series of pre-processing procedures. 1) All the letters in these documents were transformed into lower case. 2) The documents then were tokenized with non-letter separators. As a result, each document was split into a sequence of words (or tokens). 3) We removed the "stopwords" in the identified word list. "Stopwords" include the trivial English words such as "and," "the," "is," "a," "an" and so on. These stopwords don't provide meaningful information about the documents and their presence unnecessarily increases the dimensionality. 4) We removed all the tokens that are less than two letters (i.e. "s," "x," and so on), because we found those tokens don't contain meaningful information. 5) We also removed the words or tokens that appear only in one document, because these tokens are associated only with the specific study and shouldn't be considered as a reflection of any research theme. 6) We applied term stemming techniques to word list. Terms stemming will identify the root of the words and regard all words with the same root as one token. For example, "collaborate," "collaborating," "collaboration," and "collaborative" will be regarded as a single token, the "collabor–." By doing so, different variants of the same word are combined and the dimensionality is further decreased. 7) Finally, we removed "author," "paper," "conclusion" and some other words that are associated solely with the writing style of scholarly articles and don't provide additional information about the content. All these term reduction steps eventually resulted in a word list with 157 tokens for the definition data set and a word list with 3764 tokens for the abstract data set. These two word lists were processed respectively along with their original data sets in the next two steps.

Term Frequency Matrix Transformation: After the aforementioned processes, all documents are converted into a term frequency by document matrix. Each cell of the matrix records the frequency of occurrences for a particular token in particular document. Instead of using this absolute term frequency matrix directly, we transformed the values in the matrix using TF-IDF (term frequency – inverse document frequency) weighting method (Han et al. 2011; Harman 1992; Husbands Simon and Ding 2001; Salton and Buckley 1988; Salton Wong and Yang 1975). This approach puts more weight on the rare terms and discounts the weight of the common terms such as "social," "computing," so that the uniqueness rather than the commonality of each document will emerge in the result (Sidorova et al. 2008).

Singular Value Decomposition: We then applied singular value decomposition to convert the TF-IDF weighted term matrix into the production of three matrices, the term-by-factor matrix, singular value matrix (square roots of eigenvalues), and the document-by-factor matrix. The term-by-factor matrix shows the term loadings to a particular latent factor. The document-by-factor matrix shows the document loadings to a particular latent factor. The singular values (square roots of eigenvalues) represent the importance of particular factor.

Factor Interpretation

We associate each factor with its high-loading terms and documents to assist factor interpretation. For each solution, we created a table containing all high-loading terms and documents sorted by absolute loading. We then use these terms and documents to interpret and characterize (i.e., label) the factor. The process of labeling the factors consisted of examining the terms and documents (abstracts) related to a particular factor, interpreting the underlying area, and determining an appropriate label.

Measuring the Strength of Research Themes

In order to assess how different research themes change overtime, we measured the strength of a research theme as a frequency count, i.e., the number of documents that load highly on the corresponding factor. We classified each document into the particular research theme by its loadings. As discussed earlier, one document may load on multiple research themes (the cross-loadings). In this case, the document is classified to the research theme that has the highest loading. The number of documents being classified to each research theme is considered as the strength of the theme. The results of this analysis are discussed in the result section and the evolution of the research themes overtime is depicted in Figure 3.

RESULTS AND DISCUSSIONS

The Definitions for Social Computing

In order to address the posited research questions we analyze the definition dataset and the abstract dataset. For the definition dataset, we did not conduct dimension (or factor) reduction, therefore the result in Table 1 lists all the dimensions uncovered in the definition dataset regardless of the importance (the variance explained) of a dimension. LSA shows that there are 27 factors explaining all the variance in the definition dataset. The importance of each factor as indicated by the amount of variance accounted by the factor is captured by the singular values in Table 1. Each of the factors in Table 1 mathematically represents an orthogonal vector in a semantic hyperspace with 27 dimensions and therefore each factor reveals a unique component of the social computing definition found in the extant literature. The factors can also be considered as a latent construct that can be characterized by the terms which have loaded on it. Researchers don't need higher singular values in LSA, because the singular values shown in Table 1 represent the actual complexity of the text content being analyzed.

Factors	Interpretations (Labels)	Singular Values	High-Loading Terms[1]
Factor 1	An organizational Information	1.093	organ, inform, pervas, system, emerg

[1] The terms listed in this column are truncated. This is one common practice of text mining (Sidorova et al. 2008). These truncated terms or tokens were resulted from the step 6 in Term Reduction Process and helped to reduce the term list.

	Systems		
Factor 2	A service of Knowledge Extraction in Organizations	1.05	service, organ, pervas, emerg, knowledge
Factor 3	A science that researches pervasive behavior	1.004	scienc, pervas, comput_scienc, behavior, organ
Factor 4	A science that create and recreate social context and conventions	0.963	scienc, comput_scienc, convent, recreat_social_convent, human, service
Factor 5	A pervasive social experience	0.931	pervas, social_experi, experi, collect_technolog, human_social
Factor 6	A collection of technologies that extract social information	0.924	collect_technolog, extract, visual, process, social_inform
Factor 7	A collection of technologies that visualize social information	0.923	collect_technolog, visual, social_inform, digit_system, softwar
Factor 8	A development of social experience and social structure in groups or society.	0.856	develop, social_experi, experi, group, society
Factor 9	The participation of online community	0.854	particip, onlin, relationship, group, system
Factor 10	Pervasive idea extraction	0.833	pervas, extract, idea, process, user
Factor 11	Pervasive Creation and recreation of social experience.	0.81	pervas, creat_recreat, servic, form, social_experi
Factor 12	The Social interaction	0.803	social_interact, particip, user, refer, tool
Factor 13	Social Networking Sites	0.785	flickr, youtub, facebook, softwar, collect_technolog
Factor 14	Social Experience and connection within group	0.766	social_experi, phenomenon, connect, experi, group
Factor 15	Creation and recreation of the market	0.748	creat_recreat, market, organ, predict, creat
Factor 16	Computer social software	0.698	comput_social_softwar, user, social_experi, experi, web
Factor 17	Market prediction	0.692	market, predict, interact, site, creat
Factor 18	Market Research	0.679	research, market, resourc, emerg, inform_collect
Factor 19	Online social relationship	0.631	relationship, connect, web, pervas, social_relationship
Factor 20	Platform Design	0.615	design, gener, connect, techniqu, platform
Factor 21	A science of creating and recreating social convention	0.592	scienc, pervas, recreat_social_convent, creat_recreat, extract
Factor 22	Collaborative System design and programming.	0.561	system_design, collabor, team, user, program
Factor 23	Information sharing and Information exchange in teams	0.512	team, share_inform, exchang, program, platform
Factor 24	Computing Resources	0.444	resourc, comput_social_softwar, form, share_inform, softwar
Factor 25	System Design	0.425	system_design, form, collabor, design, share_inform
Factor 26	A collection of technology	0.406	collect_technolog, technolog, digit_system, pervas, form
Factor 27	Collective Action of communication and content sharing.	0.259	collect_action, commun_technolog, content_share, content, action

Table 1. The Interpretation of the Factors in Social Computing Definitions

In table 1, the 27 factors show that Social Computing is indeed a very broad and complex concept characterized using a diverse set of dimensions. This result shows that Social Computing is more than social networking sites (Factor 13). In fact, according to the singular values in table 1, social networking sites only account for 3.7% variance. It is not just an information and communication technology (ICT), (Factor 6 and Factor 7). Social computing is also a pervasive social experience (Factor 5) which occurs through the collective actions of interconnected people (Factor 27).

The Core Themes of Social Computing Research

The researchers used the abstract dataset to uncover the major research themes of social computing research. We conducted the same LSA as we did for the definition dataset without any factor or dimension reduction processing, 168 factors were revealed. These 168 factors would describe the social computing research in great detail but wouldn't provide us a big picture of the whole field. Therefore, we further conducted LSA at a much higher abstraction level. As demonstrated by Sidorova et al. (2008) LSA can be specified to give solutions at different levels of abstraction. It depends on the research objective to decide what level of abstract is needed. Because the objective of this study is to provide a big picture of the entire Social Computing literature, we need the most parsimonious LSA solution to cover the most content. However, there is no rule of thumb criterion for doing that (Sidorova et al. 2008). To get the best solution to this study, we had to trial test from the one-factor solution to the n-factor solution. We found that the 4-factor solution to n-factor solution were just the variants of the 3-factors solution. Therefore, the 3-factor solution is identified as the optimal solution for this study. A 3-factor solution of the LSA shows that the entire social computing research can be represented by three factors that are associated with the broad discipline of knowledge management. Table 2 shows that the term knowledge management appears in all three factors as a high-loading term. This implies that all three factors are addressing some knowledge management issue in a social computing context. A closer examination shows that each of these three factors actually represents a unique aspect of knowledge management. Factor 1 is more about the mining the knowledge from data with the consideration of social factors. Factor 2 is clearly more about the learning and sharing of knowledge within various entities, such as individual, groups and communities. Although Factor 3 also talks about knowledge management, its emphasis is on the content management of unstructured data such as text and images using tagging mechanisms and text mining. Therefore, the researchers label the Factor 1 as knowledge discovery, Factor 2 as Knowledge Sharing and Factor 3 as Content Management.

Factors	**Labels**	**High-loading Terms**
Factor 1	Knowledge Discovery in Social computing Context	knowledg, learn, knowledg_manag, manag, <u>intellig</u>, mine, engin, wiki, <u>data_mine</u>, social_factor
Factor 2	Knowledge Sharing in Social computing Context	knowledg, knowledg_manag, <u>wiki</u>, learn, <u>group</u>, social_factor, <u>share</u>, knowledg_manag_system, student, <u>knowledg_share</u>
Factor 3	Content Management in Social computing Context	<u>tag</u>, knowledg, knowledg_manag, manag, mine, softwar, imag, softwar_develop, <u>text</u>, <u>content</u>

Table 2. The Interpretation of the Three Factors of Social Computing Research

As discussed in previous studies (Sidorova et al. 2008), labeling factors in LSA is often a very challenging task, because in most cases, there is no corresponding phrases or short descriptions in English language that exactly match the meaning of a particular factor. Different researchers may give different labels for a particular factor by looking at the same group of high-loading terms. The meaning of each factor is compositely defined by all its high loading terms. This study doesn't attempt to argue that our labels for the three factor solution are perfect. Instead, the study attempts to better characterize the three-factor solution by providing a more detailed examination, i.e., a 10-factor solution. The 3-factor solution and the 10-factor solution share the same group of high-loading terms. But these high-loading terms are distributed over 10 factors in the 10-factor

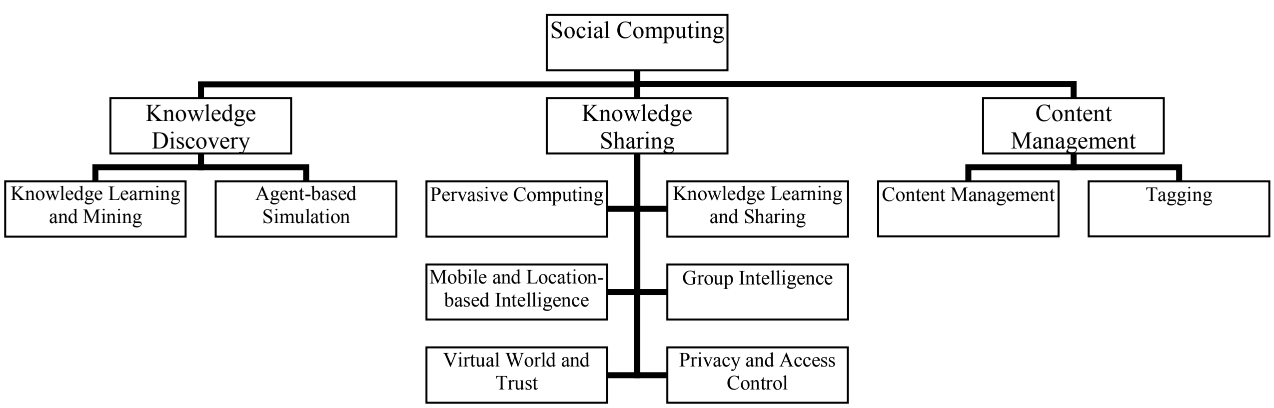

Figure 2. Social Computing Research: 3-factor solution and 10-factor solution

solution. Each factor, therefore, has less number of high-loading terms in the 10-factor solution. This means the underlying meaning of the factors in the 10-factor solution is narrower. It allows researchers to provide a more nuanced description. The researchers can also trace back these high-loading terms and examine how the factors in the 10-factors solution load to the factors in the 3-factor solution as shown in Figure 2. Finally, the researchers used the measurement discussed in the method section to assess the publication patter of each research theme in the 3-factor solution (Figure 3). Before 2007, the publications (within the context of social computing) in these three areas were sporadic and erratic. However, the results show that the research starts to converge and stabilize in these three research themes starting 2007. Among the three areas, knowledge sharing is the largest research stream which comprise of more than half of the social computing research in terms of the number of publications. While the other two themes, the knowledge discovery and content management, each account for about 20% of the total number of publications. This smoothening of the curves also indicates that the social computing discipline is starting to mature.

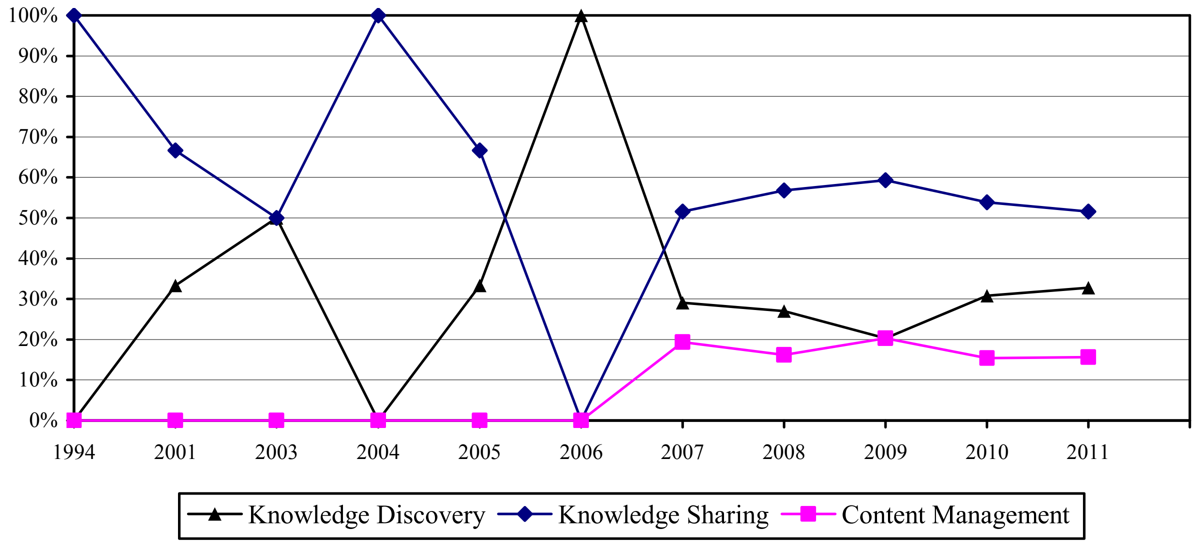

Figure 3. The Evolution of the Three Research Themes from 1994 to 2011[2]

LIMITATIONS

Although this study uses an advanced method, the Latent Semantic Analysis, to conduct a systematic and extensive literature review of publications in the social computing area, it still is subjected to some limitations which scope its interpretation and use. First, we used four databases, EBSCO, IEEEXplore, ACM Digital Library, and INFORMS, to compile our sample. Although these are the four major databases for the business and computational related publications, it is conceivable that some social computing related publications were not captured in our sample. Second, the results of this study are confined to the key words used to search and collect the papers. We only selected papers that had the whole term of "social computing" in their abstract and key words. Given that this paper focuses on the social computing phenomenon, we do not consider this as a limitation, but nonetheless it is possible that papers examining this phenomenon without explicitly using this term are omitted from our sample. Besides the limitations mentioned above, there is a special note for the readers of this paper. On one hand, LSA is a pure mathematical approach which can synthesize large body of literature in an objective manner with very little intervention from the researchers. One the other hand, LSA only provides certain statistics which requires researchers' judgment, such as in labeling of each factor. The labeling helps the researchers to make sense of the tokens uncovered from LSA. Since the underlying meaning of a factor remains unchanged regardless of what label is given to characterize the terms, we view the labeling of the terms as a means to discuss and communicate the findings of our analysis.

CONCLUSIONS

The emergence of Social Computing provides new research opportunities for IS researchers. However, this study reveals that the IS journals have not published research that focuses on Social Computing as extensively as compared to its counterparts -- IEEE (Institute of Electrical and Electronics Engineers) and ACM (Association for Computing Machinery). Among the 266 article published on Social Computing, only 26 papers are published in the IS journals. Although exploring the phenomenon

[2] No research articles were found in years from 1995 to 2000 and 2002 on Social Computing.

of SC provides new and exciting research opportunities, it also raises new challenges. The ill-defined nature of this phenomenon creates a huge challenge for the researchers. In addition, the concept of social computing is evolving because this phenomenon is still constantly growing and changing.

Unlike prior papers on Social Computing (John 2011; Parameswaran et al. 2007a, b; Wang et al. 2007; Zhou Sun Athukorala and Wijekoon 2010), where researchers' subjective judgment was used to define Social Computing and describe Social Computing research, this study makes a novel contribution by using the Latent Semantic Analysis, a mathematical natural language process technique, to synthesize all definitions and research papers on social computing to describe the current research landscape of this emerging phenomenon. The results show that there are at least 27 unique dimensions used to characterize this concept. We also conducted LSA on the abstracts from all the published social computing studies. In terms of what is Social Computing research, there is no clear boundary, yet there appears to be three dominate themes covered in the extant literature which cover multitude of context and topics. From our literature review on Social Computing, this area covers a large realm from social networking, virtual world, collective tagging, pervasive computing, and mobile computing. The research methods used in this area are also highly diverse, including network analysis, agent-based simulation, text mining, data mining, algorithm design, software prototyping and many other methods. Our data also shows that the evolution of these three themes is stabilizing over time.

REFERENCES

1. Delen, D., and Crossland, M.D. (2008) Seeding the survey and analysis of research literature with text mining, *Expert Systems with Applications*, 34, 3, 1707-1720.
2. Fox, C. (1992) Lexical analysis and stoplists, *Information retrieval: Data structures and algorithms*, 7, 102-130.
3. Han, J., Kamber, M., and Pei, J. (2011) Data mining: concepts and techniques Morgan Kaufmann.
4. Harman, D. (1992) Ranking algorithms, *Information retrieval: Data structures and algorithms*, 363-392.
5. Hossain, M.M., Prybutok, V., and Evangelopoulos, N. (2011) Causal Latent Semantic Analysis (cLSA): An Illustration, *International Business Research*, 4, 2, p38.
6. Hotho, A., Nürnberger, A., and Paaß, G. (2005) A brief survey of text mining, *Machine Learning*, 20, 1, 19-62.
7. Hristovski, D., Peterlin, B., Mitchell, J.A., and Humphrey, S.M. (2005) Using literature-based discovery to identify disease candidate genes, *International Journal of Medical Informatics*, 74, 2-4, 289-298.
8. Husbands, P., Simon, H., and Ding, C.H.Q. (2001) On the use of the singular value decomposition for text retrieval, *Computational information retrieval*, 145-156.
9. Jensen, L.J., Saric, J., and Bork, P. (2006a) Literature mining for the biologist: from information retrieval to biological discovery, *Nature reviews genetics*, 7, 2, 119-129.
10. Jensen, L.J., Saric, J., and Bork, P. (2006b) Literature mining for the biologist: from information retrieval to biological discovery, *Nat Rev Genet*, 7, 2, 119-129.
11. John, R. (2011) The Promise and Peril of Social Computing, 93-95.
12. Jungermann, F. (2009) Information extraction with rapidminer, Citeseer, 50-61.
13. Mei, Q., and Zhai, C. (2005) Discovering evolutionary theme patterns from text: an exploration of temporal text mining, in: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, ACM, Chicago, Illinois, USA, 198-207.
14. Parameswaran, M., and Whinston, A.B. (2007a) Research issues in social computing, *Journal of the Association for Information Systems*, 8, 6, 336-350.
15. Parameswaran, M., and Whinston, A.B. (2007b) Social computing: An overview, *Communications of the Association for Information Systems*, 19, 37, 762-780.
16. Salton, G., and Buckley, C. (1988) Term-weighting approaches in automatic text retrieval, *Information processing & management*, 24, 5, 513-523.
17. Salton, G., Wong, A., and Yang, C.S. (1975) A vector space model for automatic indexing, *Communications of the ACM*, 18, 11, 613-620.
18. Sidorova, A., Evangelopoulos, N., Valacich, J.S., and Ramakrishnan, T. (2008) Uncovering the intellectual core of the information systems discipline, *Mis Quarterly*, 32, 3, 467-482.
19. Turban, E., Sharda, R., Aronson, J.E., and King, D.N. (2008) Business intelligence: a managerial approach Pearson Prentice Hall.
20. Vatter, W.J. (1947) The fund theory of accounting and its implications for financial reports Arno Press.
21. Wang, F.Y., Carley, K.M., Zeng, D., and Mao, W. (2007) Social computing: From social informatics to social intelligence, *Intelligent Systems, IEEE*, 22, 2, 79-83.
22. Zhou, J., Sun, J., Athukorala, K., and Wijekoon, D. (2010) Pervasive Social Computing: augmenting five facets of human intelligence, IEEE, 1-6.